Psychotherapy and Psychosomatics

Editorial

Psychother Psychosom 2024;93:88–93
DOI: 10.1159/000535633

Received: November 3, 2023
Accepted: November 30, 2023
Published online: January 4, 2024

Rasch Analysis and Its Relevance to Psychosomatic Medicine

Kaj Sparle Christensen[a, b] Fiammetta Cosci[c, d, e] Danilo Carrozzino[f]
Tom Sensky[g]

[a]Research Unit for General Practice, Aarhus, Denmark; [b]Department of Public Health, Aarhus University, Aarhus, Denmark; [c]Department of Health Sciences, University of Florence, Florence, Italy; [d]Department of Psychiatry and Neuropsychology, Maastricht University, Maastricht, The Netherlands; [e]International Lab of Clinical Measurements, University of Florence, Florence, Italy; [f]Department of Psychology "Renzo Canestrari", University of Bologna, Bologna, Italy; [g]Department of Brain Sciences, Faculty of Medicine, Imperial College, London, UK

Keywords
Clinical assessment · Clinimetrics · Mental health · Rasch analysis · Validity

Over the years, *Psychotherapy and Psychosomatics* has published papers on numerous new psychosomatic measures [1], most notably the Diagnostic Criteria for use in Psychosomatic Research [2] and the PsychoSocial Index [3] but also measures of allostatic load [4, 5], euthymia [6], and mental pain [7]. A characteristic of all these measures is that they aim to detect and quantify clinically important concepts that cannot be measured directly. Such concepts might be termed latent traits.

These measures have all been developed using clinimetric methods and criteria [8–11]. Clinimetrics is a particular approach for clinical investigators and clinicians to develop and evaluate assessment measures intended specifically for clinical research and practice.

In the social sciences and psychology, it has been customary to develop measures starting with a large pool of items and reducing this pool to the final measure using statistical methods. This approach, part of classical test theory, is useful when researchers aim to use their data to discover new latent traits. However, problems can and do arise with the validity of traits identified in this way [12]. For example, clinicians are very familiar with the concept of burnout. The most widely used measure of burnout has been the Maslach Burnout Inventory (MBI) [13]. Some have been dissatisfied with the MBI and have argued that burnout has become the state measured by the MBI. Some studies have failed to replicate the factor structure of the MBI [14, 15]. Widespread acknowledgement that the concept of burnout had not been satisfactorily defined led to a recent Delphi study generating a consensus definition of burnout (which is rather different from that measured by the MBI) [16].

Clinimetrics started from the argument that the approach above was not only unnecessary but also inappropriate in *clinical* research and practice [17]. Rather than searching for novel concepts in their data, clinicians usually know the features of the trait they are aiming to measure [18, 19]. Items for clinimetric measures are generated not by statistical distillation of a large item pool but rather directly from clinical observation and experience [8]. Clinimetric measures include only the smallest number of items necessary to define the latent trait to be assessed. As Bech observed, clinical reality can (usually) be adequately described through a handful of items. In a robust clinimetric measure, each item refers to a feature of the underlying latent trait, which is distinct from the

karger@karger.com
www.karger.com/pps

Correspondence to:
Kaj Sparle Christensen, kasc@ph.au.dk

Karger

other items. In other words, there are no strong correlations between items – this is termed local independency. Together, the multidimensional items should characterize the trait comprehensively (in other words, there are no important features of the trait missing from the measure). To ensure structural validity, the measure should focus on an underlying trait with item scores increasing as the trait does (monotonicity), and responses to one item not affecting the responses to another (no local dependency), while ensuring that items behave consistently across various factors such as sex and age (no differential item functioning). This implies that clinimetric indices should consist of multidimensional and locally independent items. These items should consistently measure the same clinical dimension across various groups of subjects, such as men versus women or young versus old patients.

Alvan R. Feinstein, the father of clinimetrics, recommended assessing the validity of clinimetric measures on the basis of their clinical sensibility (his term) – effectively a careful appraisal that the measure was assessing what it was intended to assess [20]. Feinstein noted that clinical sensibility, which he described as "enlightened common sense," reflected a qualitative judgement. While this is a good starting point for validation, it cannot quantify the performance of a measure in relation to the features listed above. Statistical approaches are required. However, those based on classical test theory are inappropriate because clinimetric measures have not been developed according to classical test theory. For example, Cronbach's alpha is a widely used test in classical test theory of scale reliability, measuring the amount of covariance between items making up the measure, with high ratings being aimed for. As a consequence, some scales commonly used in clinical research and practice have items which are highly redundant [21] under the misguided assumption that nothing will be missed. As noted above, clinimetric measures are designed to minimize shared covariance between items as well as the number of items. Both these features will reduce item redundancy, thus Cronbach's alpha.

Rasch analysis allows measurement of the aspects of structural validity of clinimetric measures outlined above. Rasch analysis, an application of item response theory, has been widely applied to the validation of assessment measures, particularly those developed by clinimetric techniques. The main aims of this editorial are to foster further development of valid clinimetric measures by explaining the Rasch method, noting variables generated by Rasch analysis and their interpretation. This is intended to highlight validation data to be included in papers or presentations on clinimetric tools and to assist readers in appraising papers on clinimetric measures.

The Rasch model

In classical test theory, the "unit of measurement" is the total score derived from the sum of all the items of a measure. Using this assumes that every item in a measure contributes equally to the total score. This assumption is usually invalid because clinical diagnoses distinguish between major and minor symptoms but measures derived from classical test theory cannot do this [22]. For example, in the Beck Depression Inventory, the maximum score for the question on thoughts of suicide is the same as the maximum score for the question on loss of interest. In clinical practice, symptoms have different meanings in terms of quality, frequency, severity, intensity, and risks for morbidity and mortality. A person with severe suicidal ideation probably needs more urgent and intense intervention than someone with severe loss of interest. In simpler terms, scales derived from classical test theory are often considered as interval scales, implying that they not only have order but also that the gap between values holds meaning. However, in reality, they are more like ordinal scales, which rank and order data without quantifying the differences between values [22, 23].

Item response theory, by contrast, specifically considers the contributions of individual items to measuring the underlying trait and rather than assuming all items have equal weight in contributing to the underlying trait, which measures the clinical weight of each item. This supports the clinimetric principle that the independent contribution of each item of a measure should be examined because items reflecting mild, moderate, and severe symptoms need to be differentiated [8, 24].

The model developed by Georg Rasch, who was Professor of Statistics at the University of Copenhagen, is considered one of the fundamental and simplest models applying item response theory. Rasch developed his model in the early 1960s with the aim of creating a mathematical model to measure the ability of students based on their performance on a set of test items [25, 26]. Rasch was dissatisfied with the traditional approaches that relied on raw scores and treated all items as equally difficult. To address this, Rasch investigated the relationship between item difficulty, individual ability, and the probability of a correct response. He recognized that the probability of a correct response to each individual

item was dependent not only on the difficulty of the item but also on the individual's ability level. Rasch's findings led him to create a probabilistic model that estimates the probability of a correct response to an item based on the item's difficulty and the individual's ability level.

Applying Rasch analysis to psychological and psychiatric assessments, Bech and colleagues were among the first to show that someone with depression who presented with symptoms that only appear in severe states of depression (e.g., guilt feelings and psychomotor retardation) is more likely to endorse not only these symptoms but also those reflecting less severe depression (e.g., lowered mood, loss of interest, tiredness) [27]. As Bech noted [28], the hierarchical structure of items in a rating scale indicates that the underlying trait has the effect of producing unequal probabilities of item endorsement. Someone with severe depression who endorses an item regarding suicidality is also very likely to endorse items on fatigue and loss of interest, but not vice versa. Clinically, when the items of a rating scale are ordered in terms of low, medium, or high prevalence with reference to the underlying trait under evaluation, it is possible to differentiate so-called floor symptoms that emerge only in a more severe clinical condition, from ceiling symptoms which also occur in the moderate and mild forms [8, 24, 29]. In the Rasch model, the difficulty of each item is estimated separately since a measure is likely to contain combinations of items reflecting different severities of the underlying trait. The Rasch model dictates that the probability of a positive response to any particular item in a rating scale is a function of the difference between severity of the underlined trait and the characteristics of the item, namely the prevalence of the item among people showing the trait (e.g., depression). Rather than using raw item scores, the item scores are converted to the probability of an item being endorsed and the log of the probability (termed the *logit*) is used to generate a linear scale which is easier to understand visually. This has the effect of approximating the resultant scale to an interval scale, on which further analysis, including of individual items, is legitimate [30]. The Rasch model as outlined above is expressed by the following equation:

$$P(X = 1) = \frac{\exp\ (\theta - \delta)}{1 + \exp\ (\theta - \delta)}$$

where the probability $P(X = 1)$ of a respondent endorsing an item depends on the difference between the respondent's ability (θ) and the item difficulty (δ). Applied to psychological tests (e.g., patient-reported outcome measures), this corresponds to the respondent's level of the underlying trait and the item severity.

Rasch analysis assumes that an optimally designed measure will conform to the equation above. When testing a proposed measure, statistics are generated to assess the extent that the measure fits the equation, not only for the measure as a whole but also for the individual items. Given that an important principle in the development of clinimetric measures is to have the minimum number of items possible, being able to evaluate the performance of individual items is particularly useful.

Extensions of the Basic Rasch Model

The basic Rasch model is used for measurements involving items with binary responses. It primarily focuses on the distinction between item difficulty and an individual's ability or, in psychological assessments, their trait severity. The equation above can be elaborated to describe models that are more complex. The Rating Scale Model [31] is applied to measures with polytomous response categories, based on the assumption that the ordering of the response categories is the same across all items. In contrast, the Partial Credit Model [32] also allows for different response thresholds on separate items. This model extension allows for different response profiles to questions about, for example, depressed mood and questions regarding suicidal thoughts. Numerous further extensions of the Rasch model exist, but a detailed description of these extensions is beyond the scope of this editorial.

Outputs from the Rasch Model

As noted above, the key outputs of Rasch analysis all reflect the extent to which the individual items of the measure under analysis, and the overall measure itself, conform to the Rasch model expectations. In addition, the outputs help to identify how the measure could be modified to improve its fit to the model.

Summary fit statistics provide an evaluation of the fit for all items and persons. The statistics are applied to metrics which are essentially functions of the difference between observed data values and those expected from the Rasch model. The overall fit between observed and expected values can be assessed using the χ^2 statistic and standardized residuals. The latter are residuals – the difference between observed and expected value – divided by the square root of the expected value. Information-weighted fit (Infit) and Outlier-sensitive fit (Outfit) mean square statistics provide more specific details about the

Table 1. Statistical indicators for fit to the Rasch model

Fit to Rasch model	Statistical test	Satisfactory fit
Summary fit statistics	χ^2	Non-significant *p* value
	Standardized residuals of items	Mean close to 0, SD close to 1
	Standardized residuals of persons	Mean close to 0, SD close to 1
Individual item fit	χ^2	Non-significant *p* value
	Infit (unexpected responses)	Close to 1.0
	Outfit (extreme responses)	Close to 1.0
	Residuals	−2.5 to 2.5
	Standardized residuals	−2 to 2
Individual person fit	Residuals	−2.5 to 2.5
	Standardized residuals	−2 to 2

performance of the measure. Infit statistics provide information about the performance of the measure at scores close to those of most respondents, while Outfit statistics focus more on outliers.

Individual fit statistics evaluate the fit of each item or person individually to the Rasch model. These statistics focus on identifying specific items or individuals that may exhibit misfit or unexpected response patterns. Commonly used individual fit statistics include χ^2 statistic and residuals. A range of fit statistics are summarized in Table 1.

Reasons for Inadequate Fit to the Rasch Model

It is almost inevitable that data derived from a measure will not provide a perfect fit to the Rasch model. Where misfits occur, it is essential to conduct supplementary statistical assessments to identify the underlying reasons for misfit. These analyses comprise a range of strategies, including evaluation of sample size adequacy, ordering of response categories, assessment of local dependency between items, differential item functioning, targeting of the measure, and dimensionality testing of the measure.

Sample size considerations: Adequacy of the sample size is a critical factor for conducting a reliable analysis. The ideal sample size remains a subject of ongoing discussion. Research indicates the inclusion of 250–500 respondents as a reasonable guideline [33]. Alternatively, a rough rule of thumb suggests 10 times the product of the number of items and the number of thresholds [34]. For instance, in the case of an 8-item measure with four response categories (equating to three thresholds), this guideline would translate to 10*8*3 = 240 respondents. When dealing with substantial sample sizes, it is advisable to evaluate the fit to the Rasch model within adequately sized random subsets.

Differential item functioning: Ensuring consistent behaviour of a measure across different samples is crucial. Differential item functioning occurs when the measure behaves differently for distinct subgroups within the sample or for samples of different compositions. To address this, measures should be tested for differential item functioning based on factors such as, for instance, sex, age, and important comorbidities. Addressing differential item functioning is essential to mitigate bias and ensure equitable assessment. If differential item functioning is identified, there should be an explanation of how this was dealt with, such as performing separate analyses for each subgroup or deletion of items contributing to differential item functioning. Revising the wording of items contributing to differential item functioning can sometimes eliminate this problem.

Targeting of the measure: Accurate measurement depends on the items in a measure aligning well with the trait being assessed. The person-separation index is often used to quantify this alignment, which indicates how effectively the measure distinguishes individuals with varying trait levels (reliability). By using the person-separation index alongside person-item distribution maps, this alignment can be assessed visually.

A person-item distribution map visually represents how well respondents align with item difficulty. When a measure is well targeted, it can effectively differentiate respondents based on their trait levels.

Unidimensionality: The Rasch model describes a "perfect" unidimensional measure. If a measure has more than one dimension, then Rasch analysis can be applied to each dimension. If a measure fits the model, it indicates that the items collectively form a unidimensional scale

Table 2. Statistical tests to identify reasons for misfit

Identify reasons for misfit	Statistical test	Satisfactory Fit
Misfitting response categories	Compute category probability curves	Compare observed and expected probabilities
Local dependency	Residual correlations between items	Correlations <0.2 of average
Differential item functioning	Analysis of variance	No significant differences between subgroups
Multidimensionality	Principal component analysis	t test of opposing residuals less than 5%
Reliability of instrument	Person-separation index	Values >0.7

measuring the underlying trait. However, if the measure does not fit the model adequately, it may indicate multidimensionality or item redundancy. Statistical testing can be employed to assess the measure's unidimensionality more rigorously (see below).

Ordering of response categories: In a polytomous measure, a well-ordered set of response categories ensures that the choices provided to respondents reflect their level of the trait being measured. In cases with disordered thresholds, it may be necessary to collapse response categories to improve fit to the Rasch model. Adjusting the number or phrasing of disordered response categories can often resolve this issue.

Local dependency between items: Items in a clinimetric measure should be independent of each other, ensuring that they do not exhibit strong correlations. This characteristic, known as local item independence, is assessed by examining the correlations between item residuals [35]. If local item dependency is identified, it may be necessary to remove one or more of the dependent items from the scale or group them together. Parameters intended to identify reasons why a measure does not fit the Rasch model are summarized in Table 2, along with their associated statistics.

Conclusions

We have outlined key principles of Rasch analysis and highlighted some of its methods. Rasch analysis may have a place not in building a clinimetric index, but in exploring its structure and may help decreasing item redundancy. Some might question the need to gain understanding of the Rasch method. Were it not for the fact that most readers will have had thorough exposure to "classical" statistical methods starting at undergraduate level, the same argument might be applied to those methods also. Just as use of those methods depends on computer applications, Rasch analysis is facilitated by computer packages, with some dedicated to this analysis but also including Rasch modules or syntax for widely used statistical packages. Rasch techniques are not intended to replace statistics based on classical test theory. The key point is that techniques must be used appropriately. Rasch analysis has been widely used in classical psychometrics, especially for validation of measures aiming for homogeneity among items with equal weights. In the clinimetric approach, however, not all items inherently carry the same clinical weight. Importantly, the clinimetric use of Rasch analysis stands out by not relying on the psychometric assumption of homogeneity among components. Even Alvan Feinstein noted that while classical statistical methods were largely inappropriate for use in clinimetrics, these were sometimes appropriate to complement clinimetric techniques [20].

A strong case has been made, particularly in this journal, that clinimetric techniques are not only appropriate but essential to the development of measures in psychosomatic medicine [21, 36, 37]. By improving and validating measurement precision, Rasch analysis has a crucial role to play in this development.

Conflict of Interest Statement

The authors have no conflicts of interest to declare.

Funding Sources

This editorial was not supported by any sponsor or funder.

Author Contributions

Tom Sensky drafted the paper, Kaj Sparle Christensen revised all drafts of the paper with particular reference to the part related to statistics, and Fiammetta Cosci and Danilo Carrozzino revised all versions of the paper. All authors agreed with the final version of the manuscript.

References

1 Sensky T. Giovanni fava's contributions to the conceptualization and evidence base of clinimetrics. Psychother Psychosom. 2022;92(1):14–20.
2 Fava GA, Cosci F, Sonino N. Current psychosomatic practice. Psychother Psychosom. 2017;86(1):13–30.
3 Piolanti A, Offidani E, Guidi J, Gostoli S, Fava GA, Sonino N. Use of the psychosocial index: a sensitive tool in research and practice. Psychother Psychosom. 2016;85(6):337–45.
4 Fava GA, Guidi J, Semprini F, Tomba E, Sonino N. Clinical assessment of allostatic load and clinimetric criteria. Psychother Psychosom. 2010 2010;79(5):280–4.
5 Tomba E, Offidani E. A clinimetric evaluation of allostatic overload in the general population. Psychother Psychosom. 2012;81(6):378–9.
6 Carrozzino D, Svicher A, Patierno C, Berrocal C, Cosci F. The euthymia scale: a clinimetric analysis. Psychother Psychosom. 2019;88(2):119–21.
7 Fava GA, Tomba E, Brakemeier EL, Carrozzino D, Cosci F, Eöry A, et al. Mental pain as a transdiagnostic patient-reported outcome measure. Psychother Psychosom. 2019;88(6):341–9.
8 Bech P. Clinical psychometrics. Copenhagen: John Wiley & Sons; 2012.
9 Fava GA, Tomba E, Sonino N. Clinimetrics: the science of clinical measurements. Int J Clin Pract. 2012;66(1):11–5.
10 Bech P. Clinimetric dilemmas in outcome scales for mental disorders. Psychother Psychosom. 2016;85(6):323–6.
11 Carrozzino D, Patierno C, Guidi J, Berrocal Montiel C, Cao J, Charlson ME, et al. Clinimetric criteria for patient-reported outcome measures. Psychother Psychosom. 2021;90(4):222–32.
12 Fava GA, Ruini C, Rafanelli C. Psychometric theory is an obstacle to the progress of clinical research. Psychother Psychosom. 2004;73(3):145–8.
13 Maslach C, Jackson SE. The measurement of experienced burnout. J Organ Behav. 1981;2:99–113.
14 Cox T, Tisserand M, Tariz T. The conceptualization and measurement of burnout: questions and directions. Work Stress. 2005;19(3):187–91.
15 Shoman Y, Marca SC, Bianchi R, Godderis L, van der Molen HF, Guseva Canu I. Psychometric properties of burnout measures: a systematic review. Epidemiol Psychiatr Sci. 2021;30:e8.
16 Guseva Canu I, Marca SC, Dell'Oro F, Balázs Á, Bergamaschi E, Besse C, et al. Harmonized definition of occupational burnout: a systematic review, semantic analysis, and Delphi consensus in 29 countries. Scand J Work Environ Health. 2021;47(2):95–107.
17 Feinstein ART. T. Duckett jones memorial lecture. The jones criteria and the challenges of clinimetrics. Circulation. 1982;66(1):1–5.
18 Wright JG, Feinstein AR. A comparative contrast of clinimetric and psychometric methods for constructing indexes and rating scales. J Clin Epidemiol. 1992;45(11):1201–18.
19 Bech P. Modern psychometrics in clinimetrics: impact on clinical trials of antidepressants. Psychother Psychosom. 2004;73(3):134–8.
20 Feinstein AR. Clinimetrics. Yale: Yale University Press; 1987.
21 Cosci F. Clinimetric perspectives in clinical psychology and psychiatry. Psychother Psychosom. 2021;90(4):217–21.
22 Boone WJ, Noltemeyer A. Rasch analysis: a primer for school psychology researchers and practitioners. Cogent Education. 2017;4(1):1416898.
23 Wright BD, Linacre JM. Observations are always ordinal; measurements, however, must be interval. Arch Phys Med Rehabil. 1989;70(12):857–60.
24 Fava GA, Carrozzino D, Lindberg L, Tomba E. The clinimetric approach to psychological assessment: a tribute to per Bech, MD (1942–2018). Psychother Psychosom. 2018;87(6):321–6.
25 Rasch G. An item analysis which takes individual differences into account. Br J Math Stat Psychol. 1966;19(1):49–57.
26 Rasch G. Probabilistic models for some intelligence and attainment tests. Copenhagen: Danish Institute for Educational Research; 1960.
27 Bech P, Allerup P, Reisby N, Gram LF. Assessment of symptom change from improvement curves on the Hamilton depression scale in trials with antidepressants. Psychopharmacology. 1984;84(2):276–81.
28 Bech P. Rating scales for psychopathology, health status and quality of life : a comendium of documentation in accordance with the DSM-IIIR and WHO systems. Berlin: Springer-Verlag; 1993.
29 Thomas ML. The value of item response theory in clinical assessment: a review. Assessment. 2011;18(3):291–307.
30 Baylor C, Hula W, Donovan NJ, Doyle PJ, Kendall D, Yorkston K. An introduction to item response theory and Rasch models for speech-language pathologists; 2011.
31 Andrich D. A rating formulation for ordered response categories. Psychometrika. 1978;43(4):561–73.
32 Masters GN. A Rasch model for partial credit scoring. Psychometrika. 1982;47(2):149–74.
33 Hagell P, Westergren A. Sample size and statistical conclusions from tests of fit to the Rasch model according to the Rasch Unidimensional Measurement Model (RUMM) program in health outcome measurement. J Appl Meas. 2016;17(4):416–31.
34 Linacre JM. Optimizing rating scale category effectiveness. J Appl Meas. 2002;3(1):85–106.
35 Christensen KB, Makransky G, Horton M. Critical values for yen's Q(3): identification of local dependence in the Rasch model using residual correlations. Appl Psychol Meas. 2017;41(3):178–94.
36 Nierenberg AA, Sonino N. From clinical observations to clinimetrics: a tribute to alvan R. Feinstein, MD. Psychother Psychosom. 2004;73(3):131–3.
37 Fava GA. Forty years of clinimetrics. Psychother Psychosom. 2022;91(1):1–7.